Context shapes interactive alignment: the role of cumulative priming

Bert Oben & Geert Brône

Abstract

A growing body of evidence shows that dialogue involves a process of synchronisation across speakers at different semiotic levels. In this paper, we study which factors predict this synchronisation process at the lexical and gestural level. A multifactorial analysis based on a video corpus of dyadic interactions reveals that *cumulative priming* is the key factor at both levels. More than temporal or social factors, the number of preceding lexical or gestural references predicts which word or gesture participants will use. However, there is a crucial difference between the two modalities. At the lexical level cumulative priming by the interlocutor is crucial, whereas for gesture participants appear to draw on their own previous representations. A comparison with related studies shows that high-level, referential synchronisation and low-level, behavioural synchronisation seem to be governed by different rules. Models of human interaction that focus on synchronisation, should take both strands of research into account.

1. Introduction

Speakers who engage in interaction do not produce their utterances in a social-interactional vacuum, but rather do so based on and designed for an addressee. Research in various disciplines has shown that "it takes language to make language" (Du Bois 2010: 3): language use as a primarily joint activity (Clark 1996) requires speakers and their utterances to be geared to one another in multiple ways so as to facilitate fluent communication. This process of attuning crucially involves alignment[1] at different levels of linguistic organisation. Recent research in psycholinguistics, cognitive linguistics and conversation analysis has revealed that interlocutors systematically and apparently effortlessly align their linguistic representations during conversation (Pickering & Garrod 2004, 2006; Branigan et al. 2007; Menenti et al. 2012; Roche et al. 2010; Wachsmuth et al. 2013; Bazzanella 1996, Szczepek Reed 2010 and many others).

Interactive alignment as a driving force in interaction is not, however, restricted to the simple repetition of lexical items or syntactic structures in adjacent turns in conversation. Rather, recent work has shown that alignment is a dynamic contextually embedded phenomenon, in which different semiotic channels, including gesture, posture and gaze, are tightly coordinated between the

[1] We acknowledge the terminological issue that apart from *alignment* many different words refer to a comparable phenomenon: shadowing (Goldinger 1998, Lewandowski 2012), resonance (Du Bois 2010, Brône & Zima 2014), entrainment (Garrod & Anderson 1987), accommodation (Giles et al. 1992), conceptual pacts (Brennan & Clark 1996), parallelism (Tannen 1987, 1989; Sakita 2006), mimicry (Kimbara 2006), convergence (Michelas & Nguyen 2012), etc. In this paper, we use the term *alignment* as a cover term to refer to any formal features of cross-speaker repetition, regardless of any theoretical presuppositions.

interlocutors (Richardson et al. 2007, Louwerse et al. 2012, Bergmann & Kopp 2012). The present paper is intended as a contribution to the ongoing debate on interactive alignment and the mechanisms of multimodal signalling in interactional language use, and addresses two interrelated questions that have not received substantial attention in the literature:

i. which factors may explain the occurrence of interactive alignment (sequences) in longer stretches of face-to-face interaction? On the basis of different (psycholinguistic) models of dialogue, we select a series of variables pertaining to the (social) dynamics of the interaction (including speaker dominance, the temporal distance between utterances and cumulative priming) and try to model their relative impact using statistical regression analysis;
ii. Does a similar pattern of interactive alignment emerge across different modes of representation? Is gestural alignment an equally strong force in dialogue as e.g. lexical entrainment? Do the same factors predict gestural and lexical alignment?

Before addressing these questions in the empirical part of the paper (sections 5-6), using a video corpus of face-to-face interactions (section 3), we first present a brief outline of current studies on alignment, relevant for this paper (section 2).

2. Theoretical background: modelling synchronizing behaviour in dialogue

The current research interest in the phenomenon of alignment has its roots in earlier work on lexical entrainment (Brennan & Clark 1996) and referencing in dialogue (Schober 1993, Garrod & Anderson 1987), which has shown that subjects in interactional contexts tend to use the same reference terms as their interlocutors. In one line of research in psychology, these observations led to a research program on conceptual coordination processes in distributed cognition (Brennan & Clark 1996; Clark & Schaefer 1989; Clark & Wilkes-Gibbs 1986; Schober & Clark 1989). We briefly sketch the basic tenets of this program, as they are of particular relevance to the contextual features of alignment addressed in this paper.

The phenomenon of *lexical entrainment* shows that lexical choice is determined by contextual factors rather than 'ahistorical' parameters such as informativeness, availability or perceptual salience. When adopting a largely context-independent view on lexical choice, one would expect that language users determine their expression irrespective of the previous discourse sequence (and each other), but rather on the basis of conciseness and efficiency. For instance, when referring to a red car that is standing next to a van, a truck and a bicycle, speakers would tend not to use a label like *vehicle*, because it is not informative enough for referential purposes (distinguishing one vehicle from the other), or *red car*, which is too informative in the given situation. Rather, they should opt for the informatively most efficient and concise *car*. Instead of approaching the phenomenon of conceptualization and linguistic choice from such an ahistorical perspective Clark and colleagues adopt a strong interactional approach that takes partner-specific conceptualizations or *shared conceptualizations* as a driving force in dialogue. In other words, reference is argued to be designed to

a large extent with regard to the past interaction with co-participants. In a series of experiments, Brennan & Clark (1996) show that lexical choices indeed reflect the ongoing joint activity and that, hence, conceptualization in interaction is subject to a process of interactive grounding: specific sets of partners reach a temporary agreement or *conceptual pact* about a given (lexical) construal.

The interactive alignment theory (Pickering & Garrod 2004, 2005, 2006) has a different take on the parity of production and comprehension in interaction by assuming that cross-speaker alignment does not presuppose *shared* conceptualization. Rather, it is argued to be guided by a basic priming mechanism that affects the activation levels for specific (lexical) representations (Pickering & Garrod 2004: 173). On this account, (explicit) negotiation on specific representations at different levels and the modelling of others' state of mind is the exception rather than the rule in the achievement of interactive alignment: "Although we do not deny a role to intentional processes, and certainly accept that people are in principle capable of extensive modelling of their partners' mental states, we believe that the pressures of actual conversation […] mean that in practice interlocutors perform very little 'other modelling'" (Pickering & Garrod 2005: 87).

Despite the differences in the theoretical assumptions between the view of interaction as shared conceptualization and the mechanistic model of interactive alignment theory, both approaches deal with the phenomenon of routinization in interaction. Interlocutors set up local routines 'on the fly' as part of the interaction, either in the form of words or semi-fixed expressions with a conversation-specific meaning that is established and maintained during the dialogue, or as an agreement on the contextually relevant conventionalized meaning of a polysemous/ambiguous word or form. Routinization involves the setting down of new memory traces associated with a particular expression. The interactive alignment account stresses the local nature of this form of routinization (including cases of single repetition across speakers), whereas the work on entrainment seems to focus more on longer-term processes of repetition and the effect of frequency of use on the strength of local routines (Brennan & Clark 1996: 1498-1490).

A third line of research, next to the work on conceptual pacts and interactive alignment, zooms in on the analysis of matching behaviour in interaction, independent of underlying conceptual representations or situation models. More specifically, some recent studies (including Louwerse et al. 2012 and Bergmann & Kopp 2012) have explored how behaviour (including gesture, posture, and facial expression) is tightly co-ordinated between participants in an interaction. Rather than focusing on the establishment of local routines and underlying mental representations, these studies reveal the temporal organization of matching behaviour, i.e. the strong synchronisation of (non)verbal behaviours between interlocutors, independent of the conceptual content associated with specific actions. This difference between pure behaviour matching and alignment of conceptual representations will be of particular relevance to the present paper.

3. Research questions

The existing studies discussed in section 2 provide evidence for the strong force of interactive alignment in language (Pickering & Garrod 2004, 2006) and (co-speech) gesture (Bergmann & Kopp 2012, Louwerse et al. 2012), and the importance of cumulative effects in setting up dialogue-specific lexical routines (Brennan & Clark 1996). What is missing, to date, is a unified empirical account for alignment across semiotic channels, based on a single data set and analytical procedure. This leads us to the following research questions:

1. Brennan & Clark (1996) argued that lexical entrainment is subject to the frequency-of-use hypothesis, which states that "two partners should rely more on a conceptualization precedent the more firmly it has been established" (ibid.: 1498). This predication was confirmed in their study, based on a picture-naming task. What the study did not address, however, is the question whether
    a. this effect of cumulative or frequent use pertains to a co-participant's or speaker's own linguistic choices in the preceding trials. In other words, does it matter who produced the precedents in the interaction and how often?
    b. a similar effect can be measured for other than lexical references. Does the frequency-of-use effect or cumulative priming effect also appear in other semiotic channels, such as (co-speech) gesture?
    c. other factors than frequency of use may help to predict lexical or gestural alignment/entrainment.
2. Bergmann & Kopp (2012) and Louwerse et al. (2012) base their studies of non-verbal alignment on a purely behavioral approach, measuring the occurrence of gestural synchronization across speakers, independent of the conceptual representation linked to that gesture. In other words, other than in lexical studies such as Brennan & Clark (1996), they focus solely on a comparison of the physical form of adjacent gestures (e.g. hand shape, orientation etc.) and ignore the question whether these adjacent gestures (help to) express the same concept (e.g. two subsequent gestures depicting the same object). In the present study, we explore the question if
    a. taking a representational rather than a purely form-based approach to gestural alignment generates the same results as those presented in Bergmann & Kopp (2012).
    b. the occurrence of aligned gestural depictions across speakers is driven by the same explanatory principles as in lexical alignment (e.g. cumulative priming, question 1).

To address the above two sets of questions, we use a data set of interactions between two participants involved in an animation description task (section 4). The same data set will be used for both the lexical and gestural issues, so as to allow for a further comparison between the modes of representation. This comparison not only pertains to the question whether interactive alignment in the two modes is guided by the same parameters, but also whether lexical alignment typically coincides

with gestural alignment (i.e. does lexical alignment predict gestural alignment or the other way around?).

4. Data set

For this paper we use the Insight Interaction Corpus (Brône & Oben, in press), a multimodal corpus of face-to-face interactions in Dutch, transcribed and annotated for gaze and gesture. The corpus consists of conversations between 15 dyads, of about 30 minutes each. There are three subparts of the corpus: storytelling, brainstorming and targeted collaborative tasks. The results in this paper only draw on the latter part of the corpus. The collaborative tasks in the Insight Interaction Corpus are similar to the *diapix* used by Van Engen et al. (2010), in which the interlocutors play 'spot-the-difference' games on the basis of complex drawings. In that study, the interlocutors were asked to identify the differences in the drawings they were both shown (they could not see each other's pictures). The subset of collaborative tasks in the Insight Interaction Corpus differs from the diapix approach in that the animations contain moving images and the interlocutors only get to see the animation once, without being able to look back at it while discussing the differences. More precisely, the tasks in the Insight Interaction Corpus work as follows:

- two interlocutors are sitting face-to-face.
- they are each shown an animated video -simultaneously- on a screen in front of them; they can only see their own animation and not the one of their partner.
- the animations for the two interlocutors are identical except for a few details.
- immediately after seeing the animation they have to figure out the difference(s) between the videos they just saw.
- after discussing which are these differences, they are shown a new animation, and so on.

The screenshot below (fig. 1) shows the recording set-up of the Insight Interaction Corpus. On the left is the perspective of an external camera on the interaction. On the right are the videos from the mobile eye-tracking glasses the interlocutors are wearing, with the green dots indicating the visual focus. These three perspectives are edited and synchronised into one video file (from which fig. 1 is a still). For this paper we don't take into account the gaze information provided by the eye-trackers, however the scene cameras on the eye-tracking glasses do provide valuable information for a more precise view on the performed hand gestures.

Fig. 1: screenshot of recording set-up in the Insight Interaction Corpus

5. Methods

In this paper we aim at studying a number of factors that might explain why, when and how often interlocutors mimic each other's lexical and gestural behaviour. In order to be able to compare across subjects, we only take into account those lexical items and hand gestures that refer to the target objects as they appear in the animation videos. In other words, we restrict ourselves to the level of lexical representation ('which words do the interlocutors use?') and gestural depiction ('which hand gestures do they use?') of target objects.

In this methods section, we will first zoom in on the measuring technique of our dependent variable alignment (5.1). Next we discuss and motivate the choice of our independent variables, both the ones we will treat as fixed (5.2) and random (5.3) factors.

5.1 Dependant variable: alignment

5.1.1 Measuring lexical alignment

The actual data that served as the basis for our analyses can be schematically represented as in fig. 2: for each target object and per speaker, we have a lexical string of labels that were used during that conversation to refer to the target object. In fig. 2 we see the lexical string for the target object CAT for one of the conversations (see translated example below). Within those strings, we then identified the interactional pairs (indicated by the green rectangles in fig. 2). Such an interactional pair is any pair in the lexical string in which the adjacent members are uttered by different speaker. It is in those pairs that we measure whether or not the interlocutors align, i.e. whether or not they use the same label to

refer to the same target object. In the example here, there are three interactional pairs that are all aligned.

S1 First there was a **cat** and a dog.

S2 It was a black **cat**.

Did you have a black **cat** as well?

S1 Yeah.

S2 Well they started, the dog was like peeing all the time

[…]

And the **pussy** was circling, I guess it was clockwise, was circling round and round a lantern post.

S1 In my case the **pussy** was circling, the **pussy** was, I don't know, clockwise or, no I don't remember. But very fast anyway. I couldn't count how many times.

S2 The **pussy** was smaller than the dog?

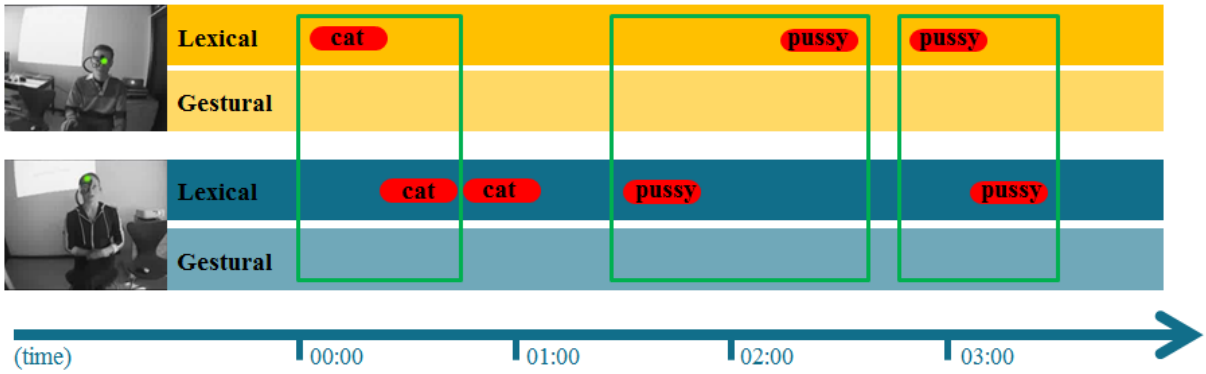

Fig. 2: lexical string for the target object CAT in which all of the interactional pairs are aligned

Scoring the dependent variable was a digital matter (there either is or there is no alignment), however, the two lexical items in the interactional pairs need not necessarily be fully identical in order to be counted as aligned. For example, we discarded diminutives and plurals and regarded cases of "katten" (cats) and "katje" (little cat) as identical and thus fully aligned to the root form "kat" (cat). Only in cases where the root forms in the interactional pair differed (like in "kat" (cat) vs. "poes" (pussy)) we considered the items in the pair as not aligned.

5.1.2 Measuring gestural alignment

When annotating and analysing gesture, it is important to take into account its innate multidimensionality. For the lexical level it is fairly easy to decide whether two cases align: there either is or there is no full alignment of the lexical root form. When annotating gesture, it is less obvious to take such digital decisions: for example, if two gestures have the same hand shape and

finger orientation but a different palm orientation, can they be considered as fully aligned? In their work on gestural alignment Bergmann & Kopp (2012) acknowledge this multidimensionality and calculate gestural alignment on one of five separate gesture features (representation technique, handedness, hand shape, palm orientation, finger orientation and wrist movement). For this study we only use one of those features to calculate gestural alignment, viz. representation technique. For the annotation of that feature we adopted the typology of depictive gestures by Streeck (2008: 292-295), who distinguishes gestural depiction methods such as modelling (hand as a token for an object), bounding (hands indicate sides or edges of an object), drawing (fingers draw lines that represent the outline or path of an object), handling (hands enact a prototypical usage of the represented object), etc.

Fig. 3: the target object DOOR is represented four times in this example

Parallel to the pairwise scoring at the lexical level, in our strings of gestural labels we scored each interactional pair referring to a target object for alignment. In the example in fig. 3 the participants are talking about a door opening in a brick wall. The target object DOOR is gesturally depicted four times in this example (red circles in fig. 3), creating two interactional pairs that are both aligned (green rectangles in fig. 4). As mentioned above, in order to label two gestures as aligned ones, for this study, we only consider the representation technique (according to Streeck 2008). This means that for the first interactional pair in the example, we measure alignment in the representation technique 'drawing', although the two gestures are not identical (the most prominent difference being that the girl uses two hands and the boy only one hand). For the second interactional pair, we see a parallel issue: the finger orientation and tension in the hand shape differ between the two speakers, but we still consider it to be an instance of gestural alignment because the representation technique is identical (i.c. modelling).

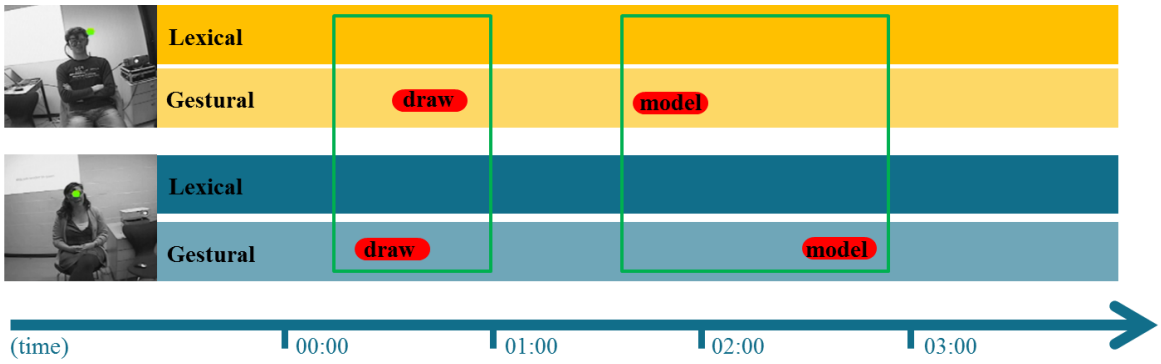

Fig. 4: gestural string for the target object DOOR *in which both interactional pairs are aligned*

5.1.3 Overcoming the content confound problem

As pointed out above, we performed a digital coding for the lexical items and gestures (plus or minus alignment). In some communicative settings, however, plus alignment cases are nearly unavoidable. Du Bois (2010: 31) refers to this issue as the *content confound* and raises the methodological question that alignment "may have simply been imposed upon the speakers by factors not entirely under their control, such as the current topic (the subject matter under discussion) and the limited set of words that the language provides for expressing this content. When two speakers engaged in conversation use the same words, isn't that just because they're talking about the same topic?".

It is plausible that if a given language only offers one lexical option to label a certain object, it is impossible for them not to align in naming that object (except for the case of circumscriptions, Costa et al. 2008). Vorwerg (2013: 152) makes the same claim, but states it the other way around: "the existence of a variety of linguistic means to express a particular idea or message both allows for and necessitates verbal attunement in communicative interaction." Du Bois (2010) uses the example of 'liver' as a referent that has no common lexical alternatives, so that if interlocutors are talking about this topic they have no option to not lexically align. However, if two speakers in two consecutive turns use the lexical item 'nerd' to refer to someone they just met, both speakers had at their disposal a vast repository of possible lexical labels to name the person they just saw (guy, dude, man, person, friend, freak, and so on) and thus had multiple options to not align. Since alignment is our dependent variable (and we are counting either plus or minus alignment cases), we wanted to rule out as much as possible the cases where the content confound makes it impossible for interlocutors not to align.

Prior to recording the Insight Interaction Corpus a pre-test was performed, in which all of the target objects in the video animations were checked for sufficient onomasiological variation potential (high lexical choice variability, cf. Brennan & Clark 1996). In a labelling game, students were asked to name a set of objects they were shown. Only if there was sufficient variation and spread of lexical labels per object, that object was selected for the video animations. That this labelling game yielded satisfactory results, will be clear from the results section.

5.1.4 Baseline comparison

As explained in the previous section, we maximally tried to avoid the content confound issue in our study. To further rule out that that we are measuring co-incidental co-occurrences of lexical items and gestures, we created a baseline comparison for our results and calculated whether there was a significant difference between that baseline and the actual results.

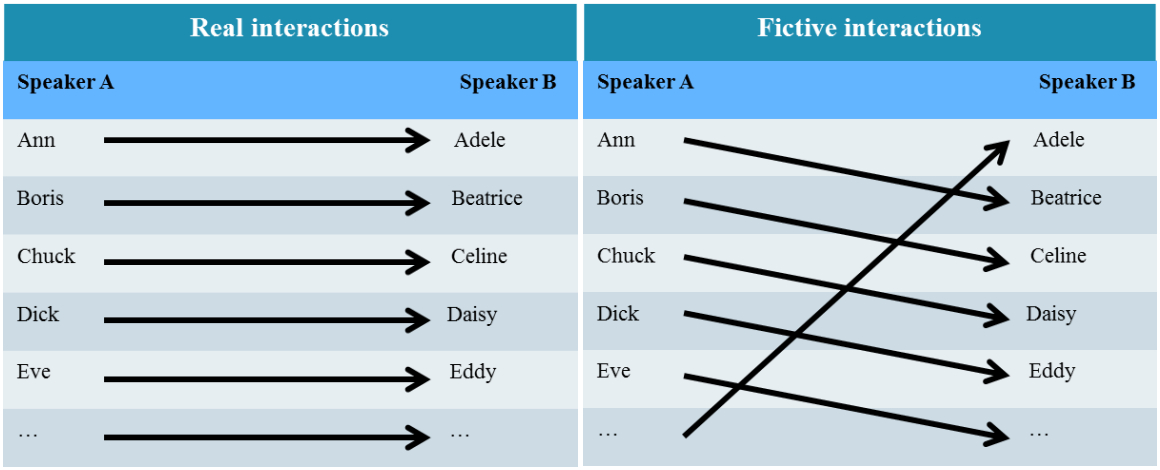

Fig. 4: the couples in the real interactions are decoupled and shuffled in the fictive interactions

The baseline in our study is a set of fictive interactions[2]. We obtained these fictive interactions by shuffling speakers so that we matched the time-aligned annotation strings of speaker A in pair 1 with that of speaker B in pair 2, speaker A in pair 2 with speaker B in pair 3, and so on (see fig. 4). Because the interlocutors in those fictive interactions are still referring to the exact same target objects, it was possible to apply the same measuring techniques for lexical and gestural alignment (as explained in sections 5.1.1 and 5.1.2). The results of measuring alignment in those fictive, shuffled interactions will form a baseline for the results of the actual conversations. To increase the reliability of this baseline we would ideally create the maximum amount of fictive interactions, i.e. to connect each speaker A with each of the speakers B from the remaining 14 interactions. However, for the scope of this study, and because the annotation process is not an automatic one, we randomized the conversational partners four times, creating 60 (fifteen dyads in the corpus that got shuffled four times) fictive interactions.

5.2 Fixed factors

So far, we have only discussed our method for measuring the dependent variable. Now we turn to the independent variables that might be good predictors for interactive alignment to occur. Each of these predictors can be linked to a specific hypothesis (for a schematic overview, see table1). It is important to note that all of the factors presented here will be used in predicting the alignment score at both the lexical and the gestural level.

[2] See Richardson & Dale (2005) or Bergmann & Kopp (2012) for a comparable baseline condition creation in conversational data.

Code	Research question *(hypothesis)*
distance	Are words/gestures closer to each other more aligned? *(y)*
position	Is there more lexical/gestural alignment towards the end of the conversations? *(y)*
animation	Is there more alignment towards the end of the experiment? *(y)*
form-self	Will speakers align more if they already used the same word/gesture themselves? *(n)*
form-other	Will speakers align more if their interlocutor already used the same word/gesture? *(y)*
concept-self	Will speakers align more if they have already referred to the same target object (but with a different word/gesture)? *(n)*
concept-other	Will speakers align more if their interlocutor has already referred to the same target object (but with a different word/gesture)? *(n)*
words	Do the most talkative speakers align the most? *(n)*
1_mention	Do the topic introducing speakers align the most? *(n)*

Table 1: overview of the fixed factors in the model

5.2.1 Temporal distance and position

The factor *distance* has already been shown to play a role in gestural alignment in an interactional setting of instruction-giving (Bergmann & Kopp 2012). For this study we calculated temporal *distance* as the time difference between (the offset of) the prime and (the onset of) the target of an interactional pair. The hypothesis is that two lexical or gestural items that occur with less elapsed time between them, have a higher chance of being aligned.

A second factor relating to the temporal dynamics of alignment is temporal *position*: for each interactional pair we calculated the relative position in the conversation[3]. Note that this is linked to the factor *animation* but the two factors should be treated separately: whereas *animation* expresses the position within the entire experiment[4], temporal *position* expresses the position within each conversation. Our hypothesis corresponds to the findings of Louwerse et al. (2012: 15): "the more interlocutors interacted with each other, the more they synchronized matching behaviors with one another"; i.e. the further into a conversation (and into the experiment), the more alignment we expect.

[3] A *conversation* refers to one of the fifteen discussions that happened after the participants saw a video animation. In a fictive example of a 280 second conversation, if the target part of an interactional prime-target pair occurred at second 140, this would be exactly halfway into the conversation, so the (relative) value for the temporal position in this case would be 0.5.

[4] This is a number between 1 and 15 that shows how many animations have already been discussed.

5.2.2 Cumulative priming

A second group of factors we included into our model concern the effect of cumulative priming. The hypothesis is that the more the interlocutors hear/see a word/gesture for a given target object, the more likely it will be that they align to that word/gesture (see Brennan & Clark 1996). To measure a possible cumulative priming effect, we combined the four following factors for each of the interactional pairs in our data set. To make this set of parameters sufficiently clear, consider the last interactional pair (rightmost green rectangle) in the example given in fig. 2.

- *form-self*: how many times, before the interactional pair, has the current speaker used the same word/gesture? (once: prior to the interactional pair the girl referred to CAT with "pussy" 1 time)
- *form-other*: how many times, before the interactional pair, has the other speaker used the same word/gesture the current speaker is using? (once: prior to the interactional pair the boy referred to CAT with "pussy" 1 time)
- *concept-self*: how many times, before the interactional pair, has the current speaker used a different word/gesture to refer to the same target object he is referring to? (twice: prior to the interactional pair the girl referred to CAT with "cat" 2 times)
- *concept-other*: how many times, before the interactional pair, has the other speaker used a different word/gesture to refer to the same target object he is referring to? (once: prior to the interactional pair the boy referred to CAT with "cat" 1 time)

With the set of factors above we not only measure whether or not cumulative priming is a significant factor in explaining alignment, it also allows us to disentangle the issue into more precise questions. Our hypothesis is that formal repetition (not the mere number of mentions of a target object) and other-priming (as opposed to self-priming) are key in predicting more lexical/gestural alignment.

5.2.3 Dominance

Social and emotional factors (Hove & Risen, 2009) have been shown to determine the occurrence and rate of alignment phenomena. Van Baaren et al. (2009: 2382) claim that speakers who are "more concerned with others, depend more on them, feel closer to them, or want to be liked by them, tend to take over their [conversational partners'] behaviour to greater extent". In line with this, Louwerse et al. (2012) show a social asymmetry of alignment in their data: in a map task experiment instruction followers imitated instruction givers significantly more often than the other way around.

Speaker dominance in this paper is measured in two ways: by the number of words uttered during the conversation (*words*) and by checking who is the first to label a given target object (*1_mention*). First, we counted the total numbers of words per interlocutor in a conversation and then

calculated the relative speaker dominance[5] within that conversation. Second, for each interactional pair in our database, we annotated who introduced the topic, i.e. who was the first to label the target object talked about. Although we acknowledge that these are very coarse measures for speaker dominance, the hypothesis is that dominant speakers (i.e. the ones talking the most and the ones that are first in referring to the target object at hand) will align less than non-dominant speakers: the latter will be more likely to 'follow' their dominant conversational partner than the other way around.

5.3 Random factors

Some *dyads* will align more than others. Also, some *target objects* will be more alignment than others. To maximally discard this variation in our dependant variable *alignment*, we will treat them as random factors in our mixed effects models.

6 Results

6.1 Baseline comparison: interlocutors align lexically and gesturally

Before turning to the analysis of the independent variables described above, we first want to demonstrate how we successfully tackled the content confound issue (see 5.1.3). In 86 % of the lexical interactional pairs (n=730) the interlocutors use the same word to refer to the same target object. Likewise, in 58 % of the gestural interactional pairs (n=543), the interlocutors use the same gestural depiction technique to refer to the same target object. In our control dataset, a set of speaker-shuffled interactions (see 5.1.4), the alignment levels are 61 % (n=1918) for lexemes and 48 % (n=1068) for gestures. The difference between the actual and the shuffled data set is significant[6] (χ^2=150.31, p<0.001 at the lexical and χ^2 =21.99, p<0.001 at the gestural level), which indicates the alignment we measure is real and not due to chance or content confound alone.

6.2 Descriptive statistics: lexical vs. gestural alignment

We have already shown that we measured more lexical (0.86 aligned pairs) than gestural alignment (0.58 aligned pairs). This frequency difference is significant (χ^2 =129.03, p<0.001). Interestingly, speakers who score high on lexical alignment do not necessarily score high on gestural alignment. Fig. 6 shows a scatter plot of the averaged alignment scores for lexemes and gestures per *speaker*. As is already clear from the plot, there is no correlation between the two (r=-0.03). Likewise, when

[5] In a fictive example of one conversation where speaker1 uses 800 words and speaker2 400 words, the relative frequencies of resp. 0.67 and 0.33 would be used as values for the independent variable *words* in our database.
[6] In this study a baseline comparison works well because the target objects talked about are controlled for: both in the actual and the shuffled data set the interlocutors are talking about the exact same things. The only thing we manipulated in the baseline is the interactionality of the data: we omitted the temporal dependencies and turned the ordered strings of references to target objects into random strings.

averaged across *target objects*, there hardly is a correlation (r=0.23): target objects that are often lexically aligned are not systematically gesturally aligned as well.

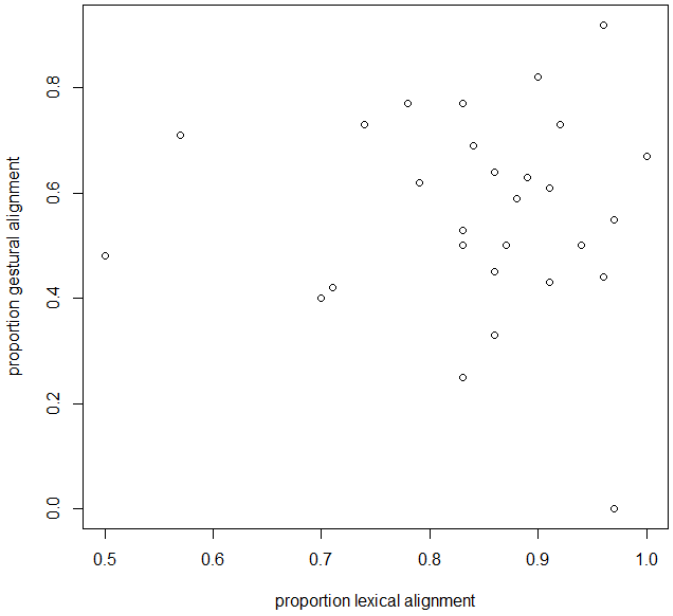

Fig. 6: crosstab of averaged lexical and gestural alignment per speaker

6.3 Mixed effects models: cumulative priming as key factor

Our baseline comparison test demonstrated there is an above chance level amount of lexical and gestural alignment in representing target objects. To show which factors predict this alignment, we used R and lme4 (Bates et al. 2014) to perform a mixed-effects regression model. The dependent variable was our alignment score per interactional pair, i.e. a binomial response variable because alignment was scored digitally (alignment: yes or no). To test for collinearity issues, we calculated Pearson correlations or Cramer's V for all possible variable interactions. None of the correlation measures (for both the lexical and gestural level) were larger than 0.30, providing sufficient evidence for the independency of our fixed factors. To determine which fixed effects (from the overview described in table 1) to enter in the model, we used a forward stepwise variable selection procedure[7]. As described in section 5.2, we used *dyads* and *target objects* as random factors in our models.

Both at the lexical (see table 2) and gestural level (see table 3) the priming factors are key in predicting alignment. However, there is a clear difference between the two levels: for lexical alignment, the accumulative behaviour of the interlocutor is crucial (*form-other* and *concept-other* in table 2), whereas for gestural alignment it is accumulated self-priming that is primordial (*form-self* in table 3). None of the other factors, and no relevant interactions between factors have been found to be significant.

[7] We used the stepAIC function in the MASS package (Venables & Ripley, 2002)

Fixed factor	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	1.909	0.267	7.150	8.69e-13	***
form-other	0.728	0.157	4.628	3.70e-06	***
form-self	0.305	0.142	2.148	0.0327	*
concept-other	-1.653	0.215	-7.699	1.37e-14	***
concept-self	-0.445	0.262	-1.699	0.0893	.

Table 2: mixed effects model for alignment at the lexical level

Fixed factor	Estimate	Std.Error	z value	Pr(>|z|)	
(Intercept)	0.018	0.232	0.078	0.9380	
form-other	0.294	0.139	2.108	0.0350	*
form-self	0.472	0.136	3.475	0.0005	***
concept-other	-0.099	0.128	-0.779	0.4360	

Table 3: Mixed effects model for alignment at the gestural level

To evaluate the predictive power of our mixed effects models we performed two tests. First, we calculated C-values for the two models. Both the lexical (C=0.89) and gestural model (C=0.75) appear to have (near to) predictive power. Second, we compared the fitted value for each data point to the actual value in the response variable[8] and found that the model predicted 90 % of the data correctly for the lexemes, and 69 % for the gestures.

7. Discussion

Existing research has shown that interlocutors match different levels of behaviour with that of their interlocutor. What separates the present study and Brennan & Clark (1996) from Louwerse et al. (2012) and Bergmann & Kopp (2012) is the measurement technique of the dependent variable *alignment* (cf. supra section 3). The former essentially deal with referential alignment (which labels do interlocutors use to refer to a given target object?), whereas the latter study behavioural alignment (which formal features of language use (including non-verbal) do interlocutors share?). In this study we used a uniform data design and method at the lexical and gestural level to uncover whether referential alignment occurs more frequently than chance (see 7.1), and to uncover by which factors it is explained (see 7.2) or not explained (see 7.3).

7.1 Referential alignment of words and gestures

Our baseline comparison test showed the alignment we measure is real and not due to chance alone. Especially at the lexical level this is an important result because the average alignment rate (0.86) is

[8] We rewrote the fitted values into a binomial dataset, with fitted values larger than 0.5 as predicting alignment (value "1"), and smaller than 0.5 predicting absence of alignment (value "0").

very high there. We successfully excluded that this high average occurs because speakers have only limited possibilities in lexically labelling the target objects. For example, even when talking about the abstract geometric object *circle*, interlocutors referred to it in many different terms such as "ball", "disc", "wheel" or "egg". In line with Brennan & Clark (1996) we observe that lexical choice variability is high between conversations, while it is relatively low within a conversation. The different speakers in our data set use many different words (and gestures) to refer to the same objects, but they tend to use the same words (and gestures) as their conversational partner.

Although lexical and gestural alignment both passed the baseline comparison test, there is an obvious frequency difference between the two levels. What might explain the fact that there is significantly more lexical than gestural alignment? A first explanatory factor is the difference in uptake between words and gestures. Both video-based as well as eye-tracking based studies (Gullberg & Kita 2009, Oben & Brône [submitted]) show that hearers fixate only a minority of speakers' gestures. If interlocutors don't see or process[9] their partner's gestures, they will align less[10]. The low level of information uptake for gestures doesn't hold for the lexical level, where nearly all of the acoustic information is processed. This difference in information uptake can partially explain why - regardless of the factors in the regression model- there is less gestural than lexical alignment. In fact, bearing in mind the very limited amount of time that interlocutors spend on focussing on each others' gestures (not more than 1% in Gullberg & Holmqvist 2006), it might even strike as surprising that gestural alignment (with 58 % of the gestures being aligned) is a prominent phenomenon at all.

A second factor contributing to the frequency difference between gestural and lexical alignment is the multidimensional nature of gesture. In our coding scheme we only took into account the gesture type, and not –among other dimensions- the place in the gesture space, finger orientation, motion path or velocity. This coarse-grained interpretation of gesture might underestimate the prominence of one of those dimensions in the process of interactive alignment. In other words, we might be measuring less formal alignment than there really is.

7.2 Key factor: Cumulative priming

If alignment (of either lexemes or gestures) were an automatic process, involving strict priming-based input-output matching, we would expect it to occur immediately from the first interactional prime-target pair, and continue ceaselessly from that point onwards. Our results, however, suggest that not

[9] It is important to point out that any study on visual fixations (with or without the help of eye-tracking tools) can only provide positive evidence: if there is a fixation on a given object, the participant has cognitively processed the visual stimulus, however, if there is no fixation, it can't be ruled out that the participant still has processed the stimulus. This is due to the human peripheral vision, which allows information uptake without explicit fixations within an angle of 120° (Duchowski 2007, p. 29-32).

[10] Our results indicate that it is interaction that causes interlocutors to align, in other words it is the processing of the other speaker's multimodal utterances that generates significantly more alignment. Turning this observation around: if this processing doesn't happen (as was the case in the fictive interactions), there is significantly less alignment. Of course, alignment is still possible although there is no processing of a previous speaker's utterance, but in those cases it is not due to interaction but to chance and content confound.

immediate, but rather cumulative priming is key for both lexical and gestural alignment. In this vein, frequency-of-use is a stronger predictor than recency. Whether or not an interactional pair is aligned, is better predicted by the number of mentions prior to that pair than by the prime in the pair alone. In other words, interlocutors (consciously or not) take into account more context than the immediately preceding utterance alone. At the lexical level, in line with Brennan & Clark (1996), it is the accumulated behaviour of the other speaker that predicts best whether or not the current speaker will align. At the gestural level, our data support the claim made by Bergmann & Kopp (2012: 1329) that "the alignment between gestures is reliably stronger within speakers than it is across speakers", making the accumulated own behaviour the best predictor. It should be noted however, there is a crucial methodological difference (see also section 7.3.1) between this study and Bergmann & Kopp: we measure referential alignment whereas they measure formal, behavioural alignment (regardless of what the gestures refer to).

Our regression analysis shows that routinization occurs, but that it should not be read as a temporal routinization (i.e. a process that takes some time), but rather as a referential routinization (i.e. a process that takes some mentions, regardless of how much time passes). The clearly non-significant factors *animation* and *position* illustrate that interlocutors do not align more as they talk longer (either throughout a single conversation or throughout the entire experiment). They only align more as they have been primed more often (by themselves or by their interlocutors). We had expected this temporal and frequency effect to coincide, but this is not the case. Apparently, the references to the target objects in our data are not evenly distributed over the conversations. More importantly, we measure no timing effect over the entire experiment. Although interlocutors grow more familiar with each other and with the video description task, they do not display more lexical or gestural alignment throughout the experiment. This observation seems to contradict the findings of Louwerse et al. (2012: 15) who claim that "the more interlocutors interacted with each other, the more they synchronized matching behaviors with one another". However, a closer look at both analyses reveals the differences reside more in the wordings than in the findings. In Louwerse et al. (2012) a temporal effect is observed in 12 out of the 19 behaviour types under scrutiny. The behaviour types concerned with the linguistic labelling of directions, colours and digits (i.e. the ones carrying the most clear propositional contents and thus the ones most related to our data set of references to target objects) did not show the temporal effect. It was the non-verbal behaviours (face and head movements) and the dialogue acts that did. With regard to the gestural level, the comparison between our study and Louwerse et al. (2012) for a timing effect is hard to make: in their case only deictic (and not iconic or symbolic) gestures passed the baseline comparison test, i.e. only deictic (and no representational) gestures were aligned significantly more often than predicted by chance.

Cumulative priming explains a lot of the alignment measured in this study[11], but as indicated by Louwerse et al. (2012: 19) there are other factors that might explain (other types of) routinization as well: "In effect, synchronization need not be *primarily* representational: it may indicate increasingly aligned perception of the external situation". Others, such as Hove & Rise (2009) or Van Baaren et al. (2009), have demonstrated how social factors are crucial in explaining alignment. In sum, referential routinization is driven by cumulative priming, as is shown in the present paper, but other types of routinization might be driven by shared communicative goals, shared physical spaces, shared emotional states, etc. and not by shared mental representations alone.

7.3 Non-significant factors

7.3.1 Distance

In our mixed effects model *distance* is not a significant factor. At the gestural level this seems to contradict the results in Bergmann & Kopp (2012) who found there is a main effect of distance between prime and target gesture. Those different outcomes can be explained by how the factor *distance* was calculated in both studies. Although they frame it as a temporal effect, for Bergmann & Kopp (2012) *distance* equals the number of gestures in between a prime-target pair. We measured *distance* in terms of seconds between the prime and target of a pair. We argue that the one cannot be taken as a proxy for the other: for example, a distance of 3 (i.e. two gestures in between prime and target, cf. fig. 6) could correspond to 5 seconds in one prime-target pair, but 2 minutes in another. When *distance* expresses the number of gestures, the factor should not be used to address temporal issues of alignment.

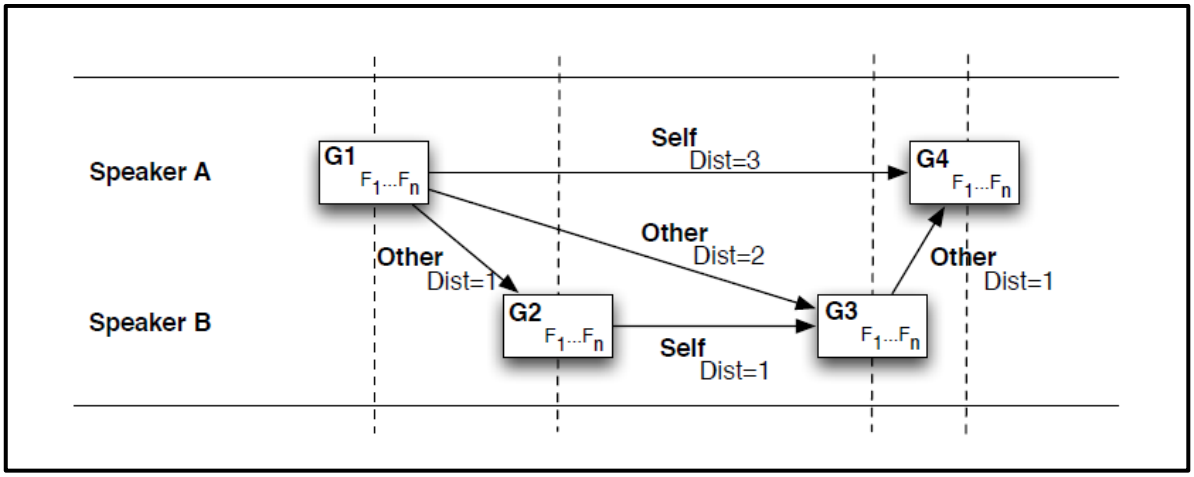

Figure 6: Schematic overview of measuring technique in Bergmann & Kopp (2012)

Parallel to the gestural level, at the lexical level our results show no significant effect for the factor *distance*. This is in line with Brennan & Clark's (1996) results, which imply "that lexical entrainment

[11] In fact, at the lexical level, cumulative priming explains more than a significant portion of the data. It explains nearly all of the data. In this vein, there is not much room left for other factors, even conversation-external factors, to account for a considerable amount of the variation in our data.

is not just a local or short-term phenomenon due to priming, but that long-term memory representations are involved." Even if prime and target are far apart, there can still be a clear alignment effect. The prominence of the factor *form-other* further illustrates that interlocutors are not exclusively primed by very recent items: a much broader context, more specifically the effect of cumulative priming, is what appears to be governing lexical alignment the most in our data.

7.3.2 Speaker dominance

The factors *words* (which speaker talks the most?) and *1_mention* (which speaker was first to introduce the target objects?) indicate that *speaker dominance* is a poor predictor for alignment. This might be due to the absence of any social hierarchy in our data. All of the participants knew each other well, they were friends and peers, and they had one common goal during the conversations, viz. to jointly try to solve the issue raised in the task ('what are the differences between the video animations for each?'). However, studies such as Louwerse et al. (2012) or Danescu-Niculescu-Mizil et al. (2012) show that if interlocutors are experimentally resp. institutionally assigned certain roles, they do show an effect of *dominance*: low power interlocutors coordinate more than high power ones. The measures for dominance in our study catch a conversation-internal type of dominance, which does not seem to generate a significant effect on language coordination as speaker *role* is shown to do. In other words, in terms of alignment frequency, it might matter more what your role is within the conversation, rather than how much you talk, or how often you introduce new topics.

8. Conclusion

There is ample evidence that interlocutors match their behaviour, both verbally and non-verbally, during interaction. When and why they do so, has only recently received substantial attention. In this paper we have shown that in referring to target objects, interlocutors in a joint task align more at the lexical than at the gestural level. Not only is there a frequency difference between those two levels, alignment is also predicted by different factors. Lexical alignment is predicted by the cumulative behaviour of the interlocutor, whereas for gestural alignment this is the cumulative behaviour of the current speaker. Moreover, there is no correlation between gestural and lexical alignment: highly aligned speakers or target objects at the one level are not systematically highly aligned at the other.

When comparing the observations concerning referential alignment in this paper to related work on behavioural alignment in other studies (most notably Louwerse et al. 2012 and Bergmann & Kopp 2012) we see that *content matters*: different factors predict different types of alignment. Behavioural alignment is predicted by speaker dominance or distance and increases as the conversation unfolds, whereas for referential alignment this does not hold true. Notwithstanding the differences between the two lines of research (behaviour matching vs. conceptual pacts), the results indicate the necessity of taking into account historical facts to account for the alignment in the data.

Ahistorical facts alone, or a fully mechanistic priming account alone, cannot account for the observations made in the growing body of research on multimodal alignment in conversation.

Acknowledgments

This research was supported by the National Fund for Scientific Research Belgium (project 3H110718: "Modelling Interactive Alignment Processes. A Multimodal and Multifocal Approach", granted to Kurt Feyaerts & Geert Brône). The authors would like to thank Kurt Feyaerts, Koen Jaspaert and Eline Zenner for their useful comments on this work.

References

Bates, D., Maechler, M., Bolker, B. & and Walker, S. (2014) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://CRAN.R-project.org/package=lme4

Bazzanella, C. (Ed.) (1996). *Repetition in Dialogue*. Tübingen: Niemeyer.

Bergmann, K., & Kopp, S. (2012). 'Gestural alignment in natural dialogue.' *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 1326 - 1331.

Branigan, H., Pickering, M., McLean, J., & Cleland, A. (2007). 'Participant role and syntactic alignment in dialogue.' *Cognition*, 104, 163-197.

Brennan, S., & Clark, H. (1996). 'Conceptual pacts and lexical choice in conversation.' *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1482–93.

Brône, G., & Oben, B. (in press). 'InSight Interaction. A multimodal and multifocal dialogue corpus.' *Language Resources and Evaluation*.

Brône, G., & Zima, E. (2014). 'Towards a dialogic construction grammar. Ad hoc routines and resonance activation.' *Cognitive Linguistics*, 25, 457-495.

Clark, H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H., & Wilkes-Gibbs, D. (1986). 'Referring as a collaborative process.' *Cognition*, 22, 1-39.

Clark, H., & Schaefer, E. (1989). 'Contributing to discourse.' *Cognitive Science*, 13, 259-294.

Costa, A., Pickering, M., & Sorace, A. (2008). 'Alignment in second language dialogue.' *Language and Cognitive Processes*, 23 (4), 528-556.

Danescu-niculescu-mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). 'Echoes of power: language effects and power differences in social interaction.' *World Wide Web: Proceedings of the 21st international conference*, 699-708.

Du Bois, J. (2010). *Towards a dialogic syntax*. Unpublished manuscript.

Duchowski, A.T. (2007). *Eye Tracking Methodology*. London: Springer.

Garrod, S., & Anderson, A. (1987). 'Saying what you mean in dialogue: A study in conceptual and semantic co-ordination.' *Cognition*, 27, 181–218.

Giles, H., Coupland, N., & Coupland, J. (1992). 'Accommodation theory: Communication, context, and consequence.' In: H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of Accommodation* (pp1-68). Cambridge: Cambridge University Press.

Goldinger, S. (1998). 'Echoes of echoes? an episodic theory of lexical access.' *Psychological Review*, 105(2), 251–279.

Gullberg, M., & Holmqvist, K. (2006). 'What speakers do and what addressees look at. Visual attention to gestures in human interaction live and on video.' *Pragmatics & Cognition*, 14, 53-82.

Gullberg M., & Kita, S. (2009). 'Attention to speech-accompanying gestures: Eye movements and information uptake.' *Journal of Nonverbal Behaviour*, 33, 251-277.

Hove, M. & Risen, J. (2009) 'It's all in the timing: Interpersonal synchrony increases affiliation.' *Social cognition* 27, 949-961.

Kimbara I. (2006). 'On gestural mimicry.' *Gesture*, 6, 39-61

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence* (Unpublished doctoral dissertation). Universität Stuttgart, Stuttgart.

Louwerse, M., Dale, R., Bard, E., & Jeuniaux, P. (2012). 'Behavior matching in multimodal communication is synchronized.' *Cognitive Science*, 2012, 36(8), 1404-1426.

Menenti, L., Pickering, M., Garrod, S. (2012). 'Towards a neural basis of interactive alignment in conversation.' *Frontiers in Human Neuroscience*, 6, 89-97.

Michelas, A., & Nguyen, N. (2012). 'Speech imitation between speakers influences the realization of initial rises in French intonation.' *Proceedings of the International Symposium on Imitation and Convergence in Speech*, 973-976.

Oben, B., & Brône, G. (submitted). 'What you see is what you do. On the relationship between gaze and gesture in multimodal alignment.' *Language and Cognition*.

Pickering, M., & Garrod, S. (2004). 'Towards a Mechanistic Psychology of Dialogue.' *Behavioural and Brain Sciences*, 27, 169-225.

Pickering, M., & Garrod, S. (2005). 'Establishing and using routines during dialogue: Implications for psychology and linguistics.' In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four Cornerstones*. London: Erlbaum, 85-101.

Pickering, M., & Garrod, S. (2006). 'Alignment as the Basis for Successful Communication.' *Research on Language and Communication*, 4, 203-288.

Richardson, D., & Dale, R. (2005). 'Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension.' *Cognitive Science*, 29, 1045-1060.

Richardson, D., Dale, R., & Kirkham, N. (2007). 'The art of conversation is coordination. Common ground and the coupling of eye movements during dialogue.' *Psychological Science*, 18, 407-413.

Roche, J., Dale, R., & Caucci, G. (2010). 'Doubling up on double meanings: Pragmatic alignment.' *Language and Cognitive Processes*, 27(1), 1-24.

Sakita, T. (2006). 'Parallelism in conversation. Resonance, schematization, and extension from the perspective of dialogic syntax and cognitive linguistics.' *Pragmatics & Cognition*, 14(3), 467-500.

Schober, M.F. (1993). 'Spatial perspective-taking in conversation.' *Cognition*, 47(1), 1-24.

Schober, M., & Clark, H. (1989). 'Understanding by addressees and overhearers.' *Cognitive Psychology*, 21, 211-232.

Szczepek Reed, B. (2010). 'Prosody and alignment: A sequential perspective.' *Cultural Studies of Science Education*, 5(4). 859-867.

Streeck, J. (2008). 'Depicting by gesture.' *Gesture*, 8(3), 285-301.

Tannen, D. (1987). 'Repetition in conversation: Toward a poetics of talk.' *Language*, 63(3), 574-605.

Tannen, D. (1989). *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse.* Cambridge: Cambridge University Press.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A.R. (2010). 'The Wildcat corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles.' *Language and speech*, 53(4), 510-540.

Van Baaren, R., Janssen, L., Chartrand, T.L. & Dijksterhuis, A.J. (2009). 'Where is the love? The social aspects of mimicry.' *Philosophical Transactions of the Royal Society B-Biological Sciences*, 364(1528), 2381-2389.

Venables, B. & Ripley, B. (2002) *Modern Applied Statistics with S.* New York: Springer.

Vorwerg, C. (2013). 'Language variation and mutual adaptation in interactive communication: Putting together psycholinguistic and sociolinguistic perspectives.' In I. Wachsmuth, J. de Ruiter, S. Kopp, & P. Jaecks (Eds.), *Advances in Interaction Studies. Alignment in Communication: Towards a New Theory of Communication*. Amsterdam: Benjamins, 149-166.

Wachsmuth, I., de Ruiter J., Jaecks, P. & Kopp, S. (Eds) (2013) *Advances in Interaction Studies. Alignment in Communication: Towards a New Theory of Communication*, Amsterdam: Benjamins, 1–10.